Stevens Institute 2021 Ensemble Flood Forecast Assessment Report: NY/NJ Harbor Area

Prepared by
Philip Orton, Ziyu Chen, Mahmoud Ayyad, Hoda El Safty
Raju Datla, Jon Miller, Muhammad Hajj

Stevens Institute of Technology

September 2022

Stevens Institute 2021 Ensemble Forecast Assessment Report: NY/NJ Harbor Area

Executive Summary[1]

The Stevens Flood Advisory System (SFAS) is a coupled hydrologic-coastal forecast system, operational since 2007 and running in an ensemble mode since 2015. Forcing from a super-ensemble of US, European and Canadian atmospheric forecast ensembles is used to reflect the uncertainty in weather forecasts. Resulting data are available on a forecast website, sent to NOAA Weather Forecast Offices to help form their Total Water Level guidance, and used for launching street-scale simulations and forecasts for Port Authority (PA) properties. The forecast system has been evaluated for winter 2016 and the year 2020, periods with generally mild or moderate storm surge conditions (up to 5 feet). Comparisons with similar NOAA forecast systems for a limited number of storms in recent years found that SFAS had better forecast accuracy and uncertainty estimates.

The primary purposes of this annual forecast assessment are to quantitatively evaluate central forecasts and uncertainty ranges, compare results to prior years, identify any problem areas and chart a path toward future improvements. We replicate the 2020 assessment metrics but also add a new assessment of the SFAS flood watch and advisory email notification system. We quantify the central forecast accuracy using RMSE, and the spread (uncertainty range) using Coverage of Observation by forecast area of Uncertainty (COU), the latter being the percentage of the time that observed water level falls within the spread (a value of 90% being optimal). As with 2020, we assess a set of five NY/NJ Harbor area stations that are representative of the harbor region's differing sub-embayments. Future years' reports (2022 onward) will expand out to regional stations.

Reviewing the year's storm surge and flood events for the Battery tide gauge, the year's peak surge was 4.1 ft, in contrast with 4.3 ft in 2020, and peak water level was 5.0 ft in contrast with 4.9 ft in 2020. The Harbor-wide RMSE on temporal maxima (peaks) for the highest five water level events of the year was 0.60 feet for lead times of 4 days, and was reduced down to 0.35 feet within one day of the peak. The RMSE for periods of time with >1.0 ft 6-hour average storm surge was 0.40 ft and for >1.5 ft surge was 0.50 ft. The spread of SFAS forecasts was very good, with Harbor-average COU of ~90.3% for cases with small surges, compared with an ideal value of 90%. Harbor-average COU averaged ~91.0% for the larger negative surges, and ~88.1% for the larger positive surges.

For Flood Watch emails, which warn of potential moderate flooding 4 days in advance, the probability of detection (POD) across harbor stations was 100% and the false alarm rate (FAR) was 94%. Given that these are based on the 95th percentile forecast exceeding the moderate flood threshold, it is unsurprising and by design that both the POD and FAR are high. For Flood Advisory emails (0-8 hours in advance), the POD was 63% and the FAR was 56% for minor floods and 100% and 98% for moderate floods. Given that the Advisories do not state what level of flooding may occur, and different locations have different thresholds for flooding, it is ambiguous how useful these results are to users. However, the SFAS website enables users to turn the stations and notifications on and off to fit their needs. Improvements could be made in the future if users are surveyed to assess their interests and needs, possibly in preparation for building a new website.

[1] For correspondence or questions, email Philip Orton at porton@stevens.edu

1. Introduction

The Stevens Institute of Technology Flood Advisory System (SFAS) is the evolution of publicly-available coastal ocean forecasting from Stevens that has grown since the inception of forecasts for the New York Harbor Observing and Prediction System (NYHOPS) in 2006 (Bruno et al. 2006) and Storm Surge Warning System in 2010 (Georgas and Blumberg, 2010). It includes rainfall-driven hydrologic inputs, tides and storm surge. Ensemble total water level forecasts have been running since 2015 and are presently based upon 96 different meteorological forecasts, providing a central-estimate time series of water level with a 90% confidence interval (Georgas et al. 2016; Jordi et al. 2019; Ayyad et al. 2022). Forecast graphics are posted with data access on the forecast webpage and interested users can sign up to be notified of impending flooding via flood watch and advisory emails.

Stevens provides probabilistic water level forecast data to the National Weather Service (NWS) Weather Forecast Offices at Upton (New York, Connecticut), Mt. Holly (Pennsylvania, New Jersey), Boston (Massachusetts, Rhode Island) and Portland (Maine, New Hampshire). NWS has been using the Stevens Flood Advisory System as an important component of their storm forecast guidance development that serves this broad region. They are now using the SFAS numeric forecast data in their Total Water Level forecast system to help inform their forecast guidance.

Stevens also provides water level observation stations and specialized forecast services to the Port Authority of New York and New Jersey (PA). These include flood simulations and forecast mapping at critical PA infrastructure sites, written forecast update reports and teleconference briefings during storm events, and flood watch and warning emails when PA sites are at risk of flooding in the future (up to 4.5 days ahead of time). These observations and forecast services help improve preparedness and resiliency at critical PA infrastructure sites, and the overall SFAS system similarly improves regional preparation and decision-making for flood events by providing unprecedented levels of accuracy and uncertainty quantification (Georgas et al. 2016; Jordi et al. 2019).

The modeling used for SFAS has been demonstrated with hindcast simulations to be capable of predicting time series or peak water levels based on atmospheric reanalysis data typically within 10-15% accuracy (RMS error), including for Hurricanes Irene (Orton et al. 2012), Sandy (Georgas et al. 2014; Orton et al. 2016; Jordi et al. 2019) and a suite of historical extreme events back to the 1700s (Orton et al. 2016). Now, the ensemble SFAS forecasts have been assessed for the winter 2015-2016 period (Georgas et al. 2016) and year 2020 (Orton et al. 2021), and here we continue with 2021 and compare results to these past assessments.

SFAS forecasts generally have lower error and more realistic uncertainty estimates in NY/NJ Harbor than NOAA forecasts. A recent direct comparison for TC Isaias (2020), the largest storm surge event since Hurricane Sandy, revealed generally better peak relative errors (PRE) and far better uncertainty estimates for SFAS versus NOAA's P-Surge. A qualitative comparison of NOAA's extratropical forecast system P-ETSS results showed it also produced considerably worse uncertainty estimates during severe storm events in 2018-2019 than SFAS did in 2020. One source of SFAS' superior uncertainty estimates is the incorporation of tides in Stevens forecast modeling, which captures tide-surge interactions that are neglected by NOAA forecasts (Ayyad et al. 2022).

The purposes of this 2021 forecast assessment and report are to:

- Archive raw model data and forecast data for future use in research
- Quantitatively evaluate central forecasts and uncertainty ranges and compare to prior years
- Continue to develop new evaluation metrics beyond those from 2021
- Identify any problem areas within our forecast system
- Chart a path toward future improvements

To focus primarily on PA interests for this report, we assess a set of five NY/NJ Harbor area stations that are representative of the harbor region's differing sub-embayments. Below, in **Section 2** we review methods behind SFAS, in **Section 3** we recap the year's significant flood and surge events and associated forecast update reports and teleconferences, in **Section 4** we report the results of our assessment, in **Section 5** we present results for Watches and Advisories, and in **Section 6** we provide some brief context, discussion and conclusions. **Section 7** also gives some future recommendations on improving the system.

2. Methods

2.1. Stevens Flood Advisory System (SFAS)

The SFAS operational hydrologic-coastal ensemble prediction system forecasts water levels across the US Mid-Atlantic and Northeast (Georgas et al. 2016). The Stevens Estuarine and Coastal Ocean Model (sECOM) is used for hydrodynamic prediction, while the US Army Corps of Engineers (USACE) packages; Hydrologic Engineering Center's – Hydrologic Modeling System (HEC-HMS), and – River Analysis System (HEC-RAS) are used for hydrologic modeling of precipitation-runoff processes and for dendritic watershed and hydraulic calculations of water flow in channels, respectively. The system is forced by various data obtained from different resources such as: United States Geological Survey (USGS), National Oceanic and Atmospheric Administration (NOAA), including 15 NY/NJ Harbor-area water level stations installed and managed by Stevens Institute.

The sECOM model is a free-surface, hydrostatic, primitive equation model with terrain-following ("sigma") vertical coordinates with an orthogonal curvilinear Arakawa C-grid. A parallelized code version is used for rapid run times and efficient use of supercomputer resources (Jordi et al. 2017). A coupled rapid surface wind-wave model helps account for wave-current combined bed stress (Georgas, 2010) and explicit representation of the effects of wave steepness on wind stress (Taylor and Yelland, 2001; Orton et al. 2012).

For the central forecast region of New York Bight, the hydrodynamic modeling is performed using a nested application of sECOM (**Figure 1**). The New York Harbor Observing and Prediction System (NYHOPS) model domain encompasses continental shelf and estuary areas with 147×452 cells, a horizontal resolution from approximately 7.5 km (4.7 miles) at the open ocean boundary to less than 50 m (164 feet) in NY/NJ Harbor, and 10 vertical sigma layers. The NYHOPS domain simulations are run in sECOM's three-dimensional mode and boundary conditions are applied at its offshore boundary (OBCs) and its interface with hydrologic models. The OBCs are a sum of (a) storm surge modeled on the Stevens Northwest Atlantic Prediction (SNAP) domain, (b) tides from the ADCIRC East Coast tide constituent database (Mukai et al. 2002), (c) a uniform 11 to 13 cm (~0.5

ft) cross-shore slope positive toward land (Georgas and Blumberg, 2010), and (d) a bias correction based on coastal tide gauge stations. Simulations span 108 h from present and are repeated every 6 hour with “00z”, “06z”, “12z” and “18z” (GMT times in hours) launch times.

SFAS uses an ensemble forecasting approach with sECOM and hydrologic simulations on each domain run with 96 different atmospheric forcing datasets. The ensemble meteorological forcing datasets are shown in **Table 1** and include the Global Forecast System and Global Ensemble Forecast System (GFS, GEFS; Han and Pan, 2011), North American Mesoscale forecast system (NAM; Skamarock et al. 2005), Canadian Meteorological Center (CMC) global ensemble prediction system (Charron et al. 2010), European Centre for Medium-Range Weather Forecasts (ECMWF; Buizza et al. 2007; ECMWF, 2018) ensemble and high-resolution member (ECMWF-HRES). Meteorological data are spatially and temporally interpolated using bicubic and cubic spline interpolation, respectively, to create hourly forcing fields on each domain’s grid.

Two ensemble forecasts are produced within the SFAS system, the regional NYHOPS-E forecasts, which include tidal forcing in the hydrodynamic simulations, and NW Atlantic SNAP-Ex forecasts, for which the tides are added after the simulations are completed. Only NYHOPS-E forecasts are evaluated in this forecast assessment, given that SNAP-Ex forecasts are not available for harbor areas. Additional nested subgrid modeling for the 5th, 50th and 95th percentile scenarios is performed for PA critical infrastructure sites with resolutions ranging from 3 to 10 meters. These results are mapped for PA properties in the (private) Port Authority Flood Advisory System (PAFAS).

Table 1: Information on meteorological forcing datasets

Product	**# members**	**Spatial resolution (°latitude)**	**Temporal resolution (h)**
GFS	1	0.25	3
GEFS	21	0.5	3
CMC (GEPS)	21	0.5	6
NAM	1	0.1	3
ECMWF-ENS	51	0.25	3
ECMWF-HRES	1	0.125	3

After every cycle of model simulations, the SFAS system splits the data into two periods; one hindcast day used for bias correction and weighting purposes and 4.5 forecasting days using the ensemble processing methods below. Hydrologic and hydrodynamic simulations, ensemble analyses and website graphics require 2 hours, and posting of time series water level forecasts occurs at about 2, 8, 14, 20h GMT (3, 9, 15 and 21h Eastern

Standard Time) at http://www.stevens.edu/SFAS. Results are then used for the PA subgrid simulations and posted to PAFAS about one hour later.

All model simulations and resulting SFAS and PAFAS forecasts are run at Stevens Institute in the Pharos Hyperscale Computing Facility, a 1,320-core Hewlett-Packard HPC built using HP Proliant servers, Mellanox FDR InfiniBand network and Seagate 2.2 PB Lustre-based storage facility, incorporating special-purpose modeling, database, and web presentation systems. The systems are housed in a custom data center designed to provide a high level of redundant power and cooling capacity to ensure uninterrupted operations.

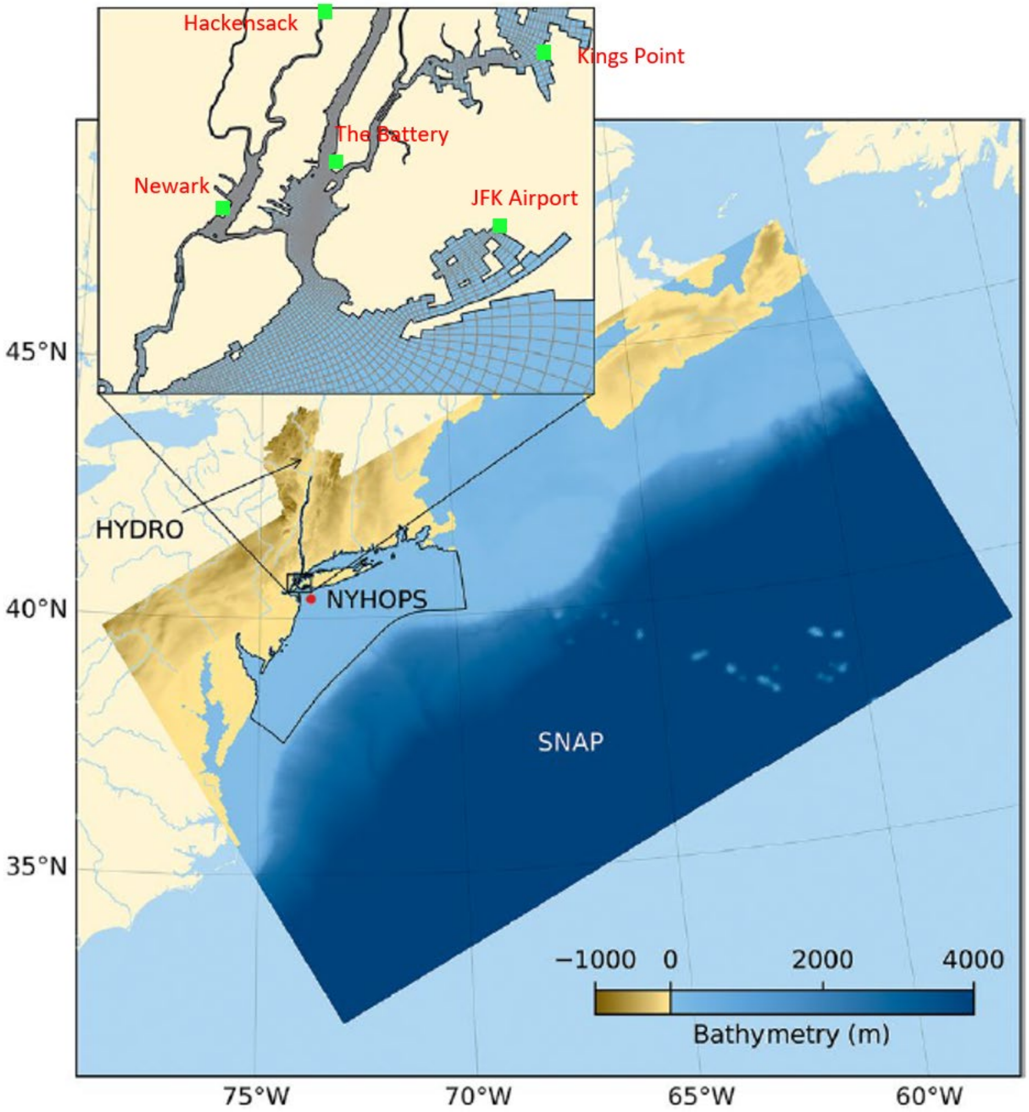

Figure 1: NYHOPS model domain linked to offshore SNAP model and inland Stevens HYDRO models. SFAS forecast locations assessed in this report include The Battery, Newark Bay ("Elizabeth Channel near Newark NJ" in SFAS), JFK Airport ("Bergen Basin at Jamaica Bay"), Kings Point and Hackensack River at Hackensack NJ.

2.1.2. Ensemble processing methods

SFAS uses a weighted average of the 96 members to find the central forecast (Georgas et al. 2016):

$$\eta_w = \sum_{j=1}^{m} w^{(j)}\eta^{(j)} \qquad (1)$$

Here, the superscript j denotes the ensemble number, m is the total number of ensembles, and w is the normalized weight which is defined by:

$$w^{(j)} = \frac{factor^{(j)}}{\sum_{j=1}^{m} factor^{(j)}} \qquad (2)$$

where factor(j) is the weight value:

$$factor^{(j)} = \frac{1}{\left(\left|\epsilon^{(j)}\right| + 0.05\right)\left(RMSE^{(j)} + 0.05\right)} \qquad (3)$$

Here, $\epsilon^{(j)}$ is the 24-hour hindcast mean bias and $RMSE^{(j)}$ the root mean squared error for member *j*.

$$RMSE^{(j)} = \sqrt{\frac{1}{N^{(j)}}\sum_{i=1}^{N^{(j)}}\left(\eta_{m\ i}^{(j)} - \eta_{oi}\right)^2} \qquad (4)$$

Thus, the normalized weights are estimated posterior model probabilities (Georgas et al. 2016).

The 5th and 95th percentiles are interpolated using the empirical distributions formed with all available ensemble members (equally weighted). The range of the ensemble represents atmospheric uncertainty across multiple ensembles, but does not include any ocean modeling uncertainty. As a rough method for approximating the effect of additional ocean, wave and air-sea interaction uncertainties, the central forecast RMSE from the hindcast period is added to the 95th percentile and subtracted from the 5th percentile to broaden the spread.

2.2. Model performance quantification

In this report, focusing on PA interests, we assess forecasts from SFAS NY/NJ Harbor-area stations for all available archived data in 2021 (all stormy periods and most of the year's non-storm periods). Five stations are representative of the major harbor embayments and tributaries -- The Battery (Manhattan), Newark Bay ("Elizabeth Channel near Newark NJ" in SFAS), JFK Airport ("Bergen Basin at Jamaica Bay"), Kings Point and Hackensack River at Hackensack NJ (**Figure 1**). A subsequent paper for peer-reviewed publication will be prepared in fall 2021 that includes stations across the NYHOPS domain. Observation stations for these five sites are run by NOAA, Stevens, Stevens, NOAA and USGS, respectively.

In this study, we adopt the RMSE, and the Coverage of Observation by forecast area of Uncertainty (COU) as our performance metrics to measure the performance of the numerical model central forecast and uncertainty range with respect to the available observation,

RMSE is here computed in four ways: (a) As shown in Eq (4), for each 5.5-day simulation (hindast and forecast period) of every ensemble member, (b) on the year's top-5 daily temporal maxima, (c) using all data within 6-hour bins of lead time for ensemble-based central forecasts, and (d) for ensemble-based central forecasts for each 6-hour forecast period separately (the "6-hour RMSE"), then presented as averages and 90% variability ranges. The latter computation used 50% overlapping bins, and for these the 6-hour average surge is also computed, to optimally capture cases with high positive or negative surges and avoid dividing one event into two bins.

COU is computed as:

$$COU = \frac{n}{N} * 100\% \qquad (5)$$

Here, n is the number of cases where the observation falls within the forecast 90% confidence interval, and N is the total number of forecast time periods. Given that there are 4 forecasts per day, each with 108 hours, each with data every 10 minutes (6 data points), thus in a full year, N = 946080 (if there are no gaps in the observational data). Again, 6-hour, 50% overlapping bins were used for COU computations.

3. Recap of 2021 Storm Events and Forecast Reporting

Significant water level and storm surge events are summarized in **Table 2** and **Table 3** below. The year's highest water level occurred during the severe nor'easter in February (**Table 2**). Three events caused an exceedance of the National Weather Service "minor flood" threshold of 4.42 ft NAVD88, and none the moderate flood threshold of 5.72 ft. The year's top water levels were observed during either severe extratropical cyclones or king tides with small storm surges of only about one foot.

The year's peak 6-hour average for storm surge occurred during a storm in October, at 3.28 ft, whereas the highest water level during nor'easter in February took second at 2.93 ft (**Table 3**). The year's highest storm surge without temporal averaging at The Battery occurred during the storm in October at 4.11 ft, which is also the largest surge since Hurricane Sandy and Isaias. Fortunately, it occurred near low tide and total water level was not in the top-5 (**Table 2**).

SFAS and PAFAS forecasts ran and the website reported the forecasts (as always) throughout the year, although there were no major flood events. An example of the online presentation of the SFAS forecasts is shown in **Figure 2**, demonstrating the forecast prior to passage of Tropical Storm Henri along with the perspective in hindcast. Observations (red dots) typically fall within the forecast 90% confidence interval (grey shading) but can deviate further from the central forecast (magenta line) in extreme cases.

Table 2: The top 5 peak water level events at The Battery station

Rank	Date	Peak water level (ft NAVD88)	6h-avg surge (ft)	Peak surge (ft)	Notes
1	2/2/2021	5.01 (minor[a])	2.93	3.33	"February severe nor'easter"
2	5/29/2021	4.80 (minor[a])	1.58	1.79	
3	10/9/2021	4.52 (minor[a])	1.48	1.83	
4	1/6/2021	4.37	1.17	1.53	
5	3/29/2021	4.23	1.13	1.03	

a: The National Weather Service minor flood threshold (4.42 ft) was exceeded at The Battery, indicating minimal or no public property damage (e.g. shallow street flooding)

Table 3: The top 5 peak events for 6-hour average surge at The Battery station

Rank	Date	6h-avg surge (feet)	Peak surge (feet)	Peak water level (ft NAVD88)	Notes
1	10/30/2021	3.28	4.11	4.03	
2	2/2/2021	2.93	3.33	5.01 (minor[a])	February severe nor'easter
3	5/29/2021	1.58	1.79	4.80 (minor[a])	
4	10/9/2021	1.48	1.83	4.52 (minor[a])	
5	9/2/2021	1.45	2.10	2.92	post-Tropical Cyclone Ida

A small subset of potential flood events prompted detailed forecast reporting and teleconference briefings by the Stevens team, including tropical storm Elsa on July 9 (5 reports, 1 conference call), tropical storm Henri on August 22 (7 reports, 2 calls), the "February severe nor'easter" on February 2 (7 reports; 4 calls), and two other storms in February (2 reports, 2 calls).

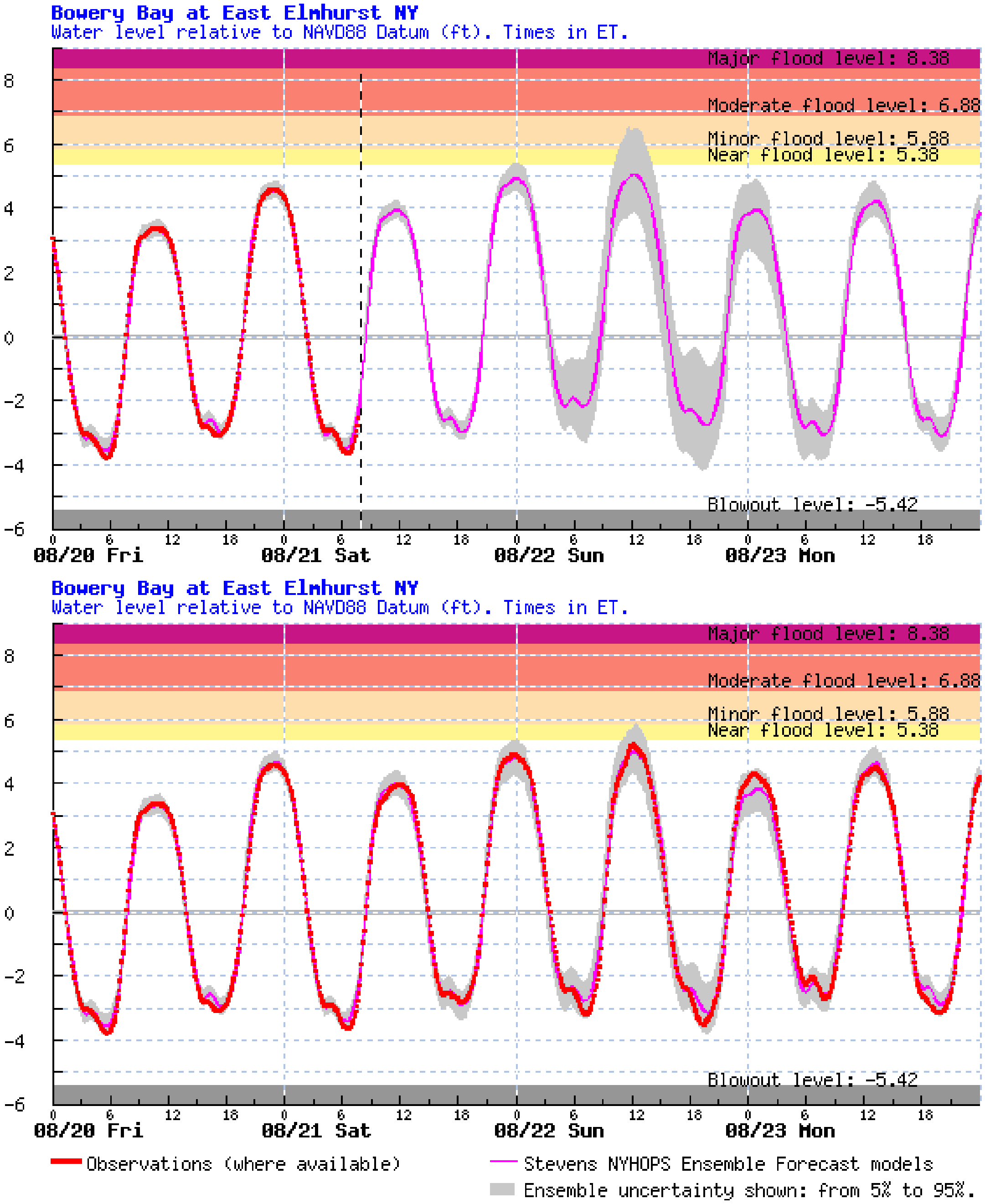

Figure 2: SFAS forecast time series of water level (top) before, and (bottom) after the passage of Tropical Storm Henri, showing good accuracy a few days in advance. Predicted peak water level was 5.0 ft NAVD88, and the observed water level came in at 5.2 ft (a -4% error). Times shown are Eastern Standard Time (ET).

4. Results: Detailed Forecast Assessment

First, we review the forecast datasets being evaluated in this report, as well as any dropouts of ensemble members through the year. **Figure 3** depicts the time series RMSE at The Battery for all members and simulations across four forecast cycles a day. The Battery observational dataset is nearly complete (no gaps), so this station and figure is useful for observing cases where ensemble members failed, sometimes regularly, sometimes episodically.

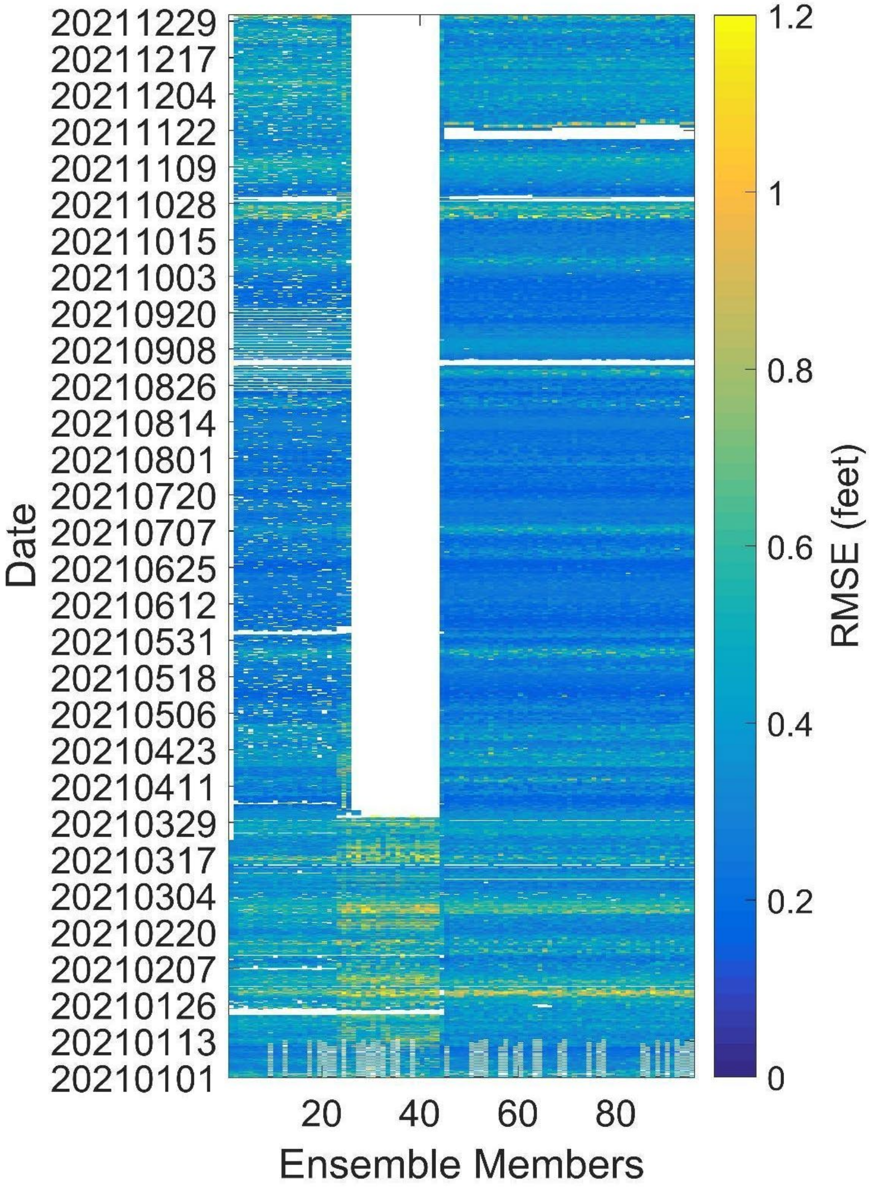

Figure 3: Assessment for individual ensemble members at The Battery station for 2021, after tide bias and mean bias corrections have been applied. The color shading is the RMSE of each member's forecast water level (relative to observed water level) across the simulation's full time range. The Y-axis is the forecast launch time, and the X-axis is the number for each of the NYHOPS Ensemble Members (**Table 1**).

The white space column denotes no simulations, as most of the CMC members 23-34 were stopped due saving computation power. Remaining white spaces or sporadic white flecks denote missing model data, which typically arose due to forcing data availability problems (e.g. with NOAA's NOMADS website) hardware failures (e.g. a computational node going down), or at times simple model or code crashes.

Typical RMSE for ensemble members is 0.2-0.4 ft, though the CMC members (#20-43) often have noticeably higher values than others. Some periods have higher RMSE values across all members. For example, in early February 2021, when an extratropical cyclone passed nearby and caused a rapid large storm surge, RMSE values were approximately 0.6-0.8 ft.

Overall, the figure demonstrates that there were at least 50 ensemble members through nearly the entire year, more than enough to form a probability distribution and estimate the 5th, 50th and 95th percentiles. Three exceptions with well below 50 members, visible as nearly complete horizontal bands of white in **Figure 3**. Two cases were September ~3 and October 29, where we see no members or very few members. The September problems arose from model crashes, identified as arising due to extreme streamflows after post-tropical cyclone Ida caused heavy rains. These are known to arise in the Raritan River basin in the NYHOPS model and caps should be placed on streamflows from the Raritan to avoid this problem, however rare (It last occurred for simulations of Tropical Storm Irene in 2011). The late October case occurred for unknown reasons, and could simply be a missing backup of forecast data. The third case began on November 20, when all the ECWMF members failed for four forecast cycles.

Focusing on the central forecast accuracy during the year's largest events, **Figure 4** shows the RMSE computed on peaks, across all five harbor stations. The harbor-wide RMSE on temporal maxima (peaks) for the highest five water level events of the year was 0.60 feet for lead times of 4 days, and was reduced down to 0.35 feet within one day of the peak.

Evaluating the forecasts at all times through the year, not just peaks during the highest water level events, the year-round forecast accuracy versus lead time is shown in **Figure 5**, computed with RMSE. All cases in **Figure 5** show improving accuracy as the date of the forecast water level gets closer (as lead time shrinks). Best performance, based on ordering of RMSE results, is Battery, Newark Bay, Kings Point, JFK Airport, then Hackensack River. RMSEs were reduced by about 20-35% from lead times of 105 hours down to 3 hours (e.g. from 0.30 to 0.22 ft at Battery, a 35% reduction).

The 6-hour RMSE mean and its 90% variability range versus lead time at each station is shown in **Figure 6**. While computed slightly differently (see detailed methods in **Section 2.3**), the mean of 6-hour RMSE shown here are very similar to the RMSE values shown in **Figure 5**. However, the addition of the variability range in this figure illustrates that RMSE for any 6-hour period can vary significantly from the mean. Next, we examine how this variability relates to storm surge.

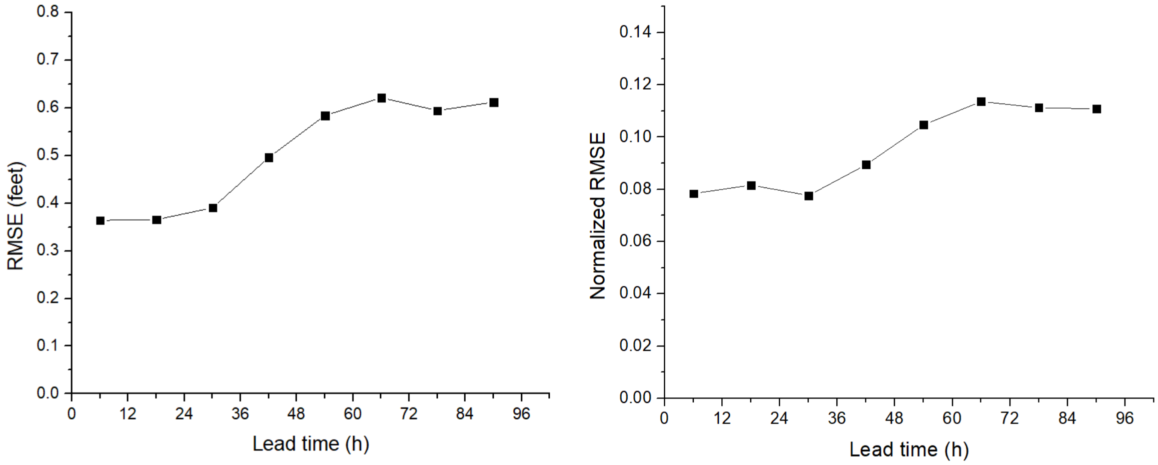

Figure 4: Assessment of accuracy of central forecasts for temporal maxima (peaks) for the top-5 events of each harbor station with RMSE.

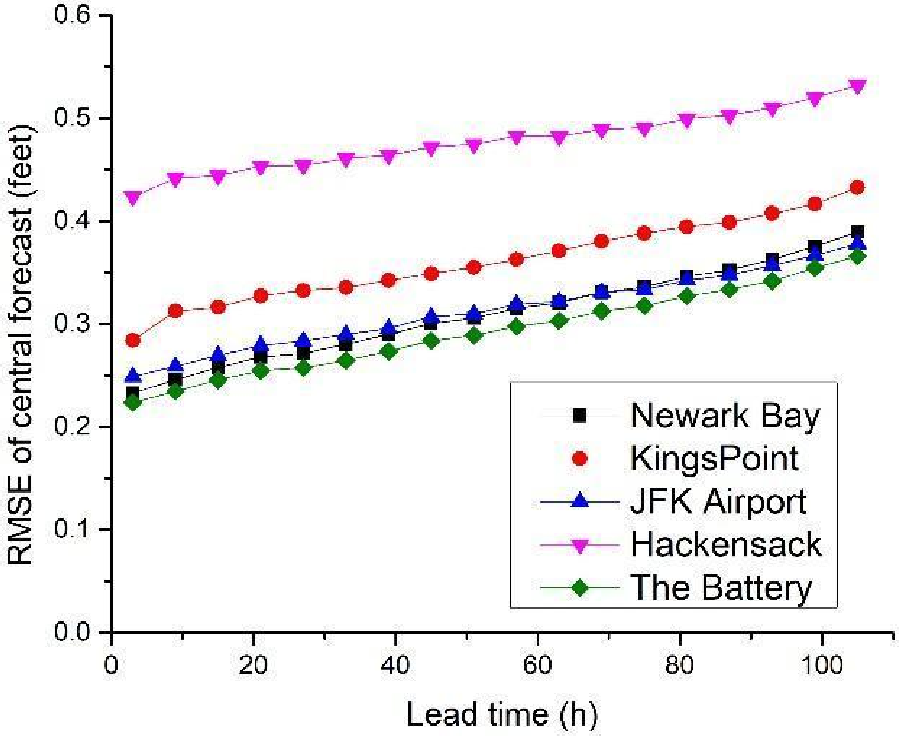

Figure 5: The SFAS weighted mean ensemble total water level forecasts against observations across the five representative harbor stations. RMSE of the weighted mean against observed water level across the forecast time in 2021. The X-axis is the forecast lead hours.

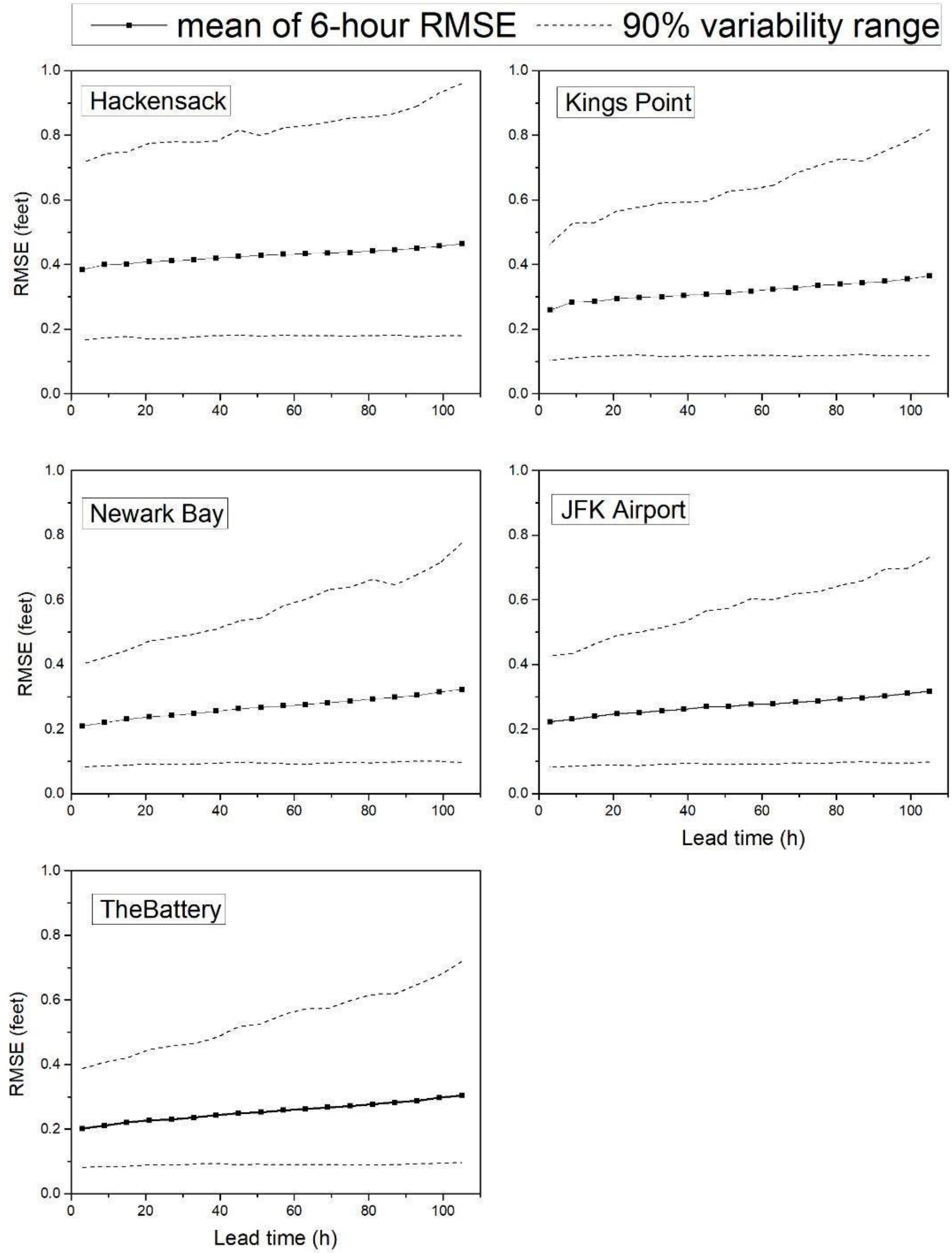

Figure 6: The weighted mean against observations across various forecast lead times and its variability range for the harbor stations. The solid lines with points are the mean of the RMSE of the weighted average against observed water level for various forecast lead hour time ranges (e.g. 0-6 hours) for 2021. The dashed lines are their corresponding 90% variability range.

The accuracy of central forecasts is particularly important during storm events. It is assessed in **Figure 7** where the mean 6-hour RMSE and its variability with storm surge is plotted for each station. Typically RMSEs grow with surge amplitude (positive or negative). At Battery, for example, the RMSE for cases with >1.0 ft surge is 0.47 ft (90% variability from 0.13- 0.80 ft) whereas for cases with >1.5 ft surge it is 0.62 ft (0.17-1.05 ft). Typical total water levels are 4.5-5.0 feet at Battery in these cases (**Table 3**), and therefore the relative error averages 12-15% during the year's worst storm surge events.

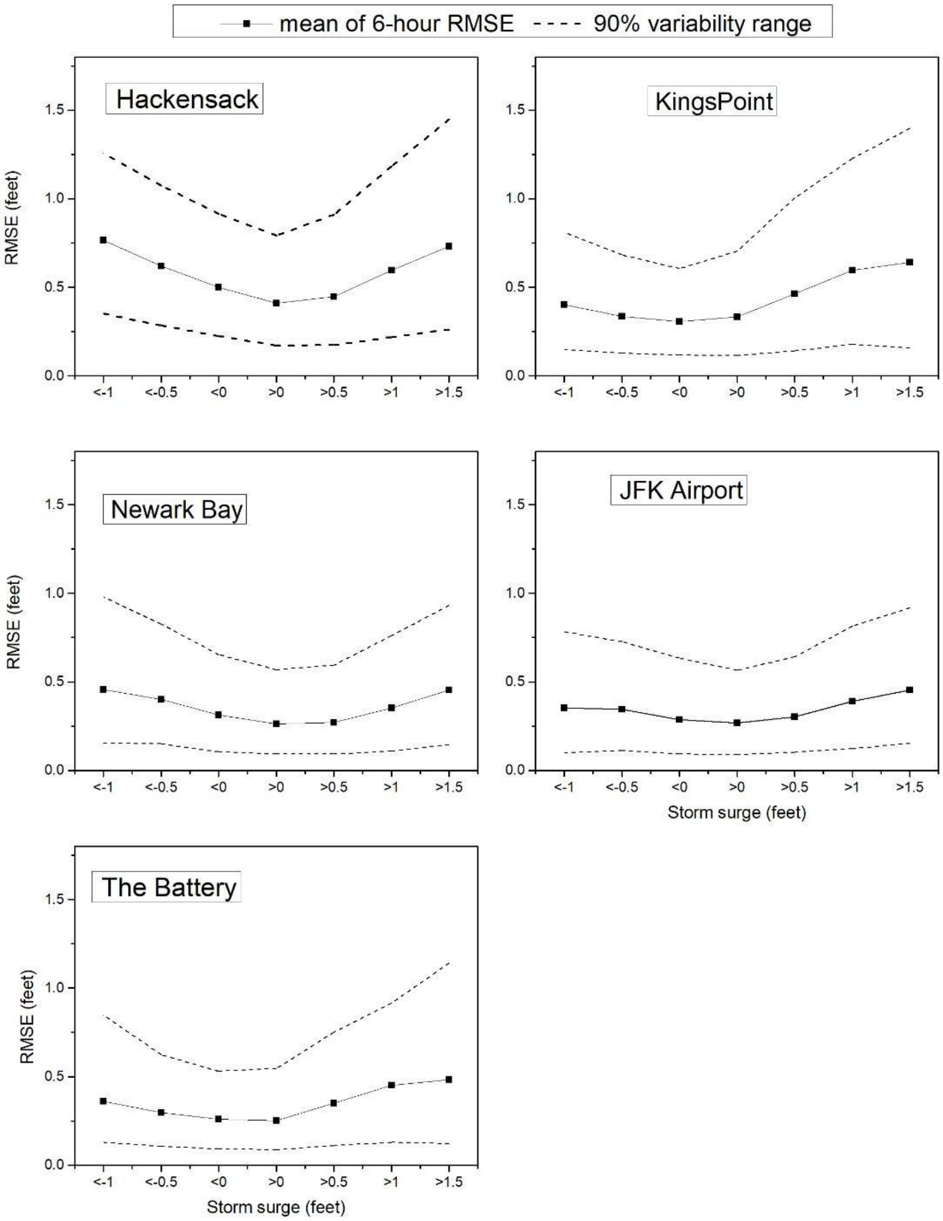

Figure 7: The RMSE of the weighted mean for various surge thresholds for the four harbor stations. RMSE and mean surge are computed in 6-hour periods, the black squares are the mean of RMSE results across all lead times and the dashed lines are the range within which 90 percent of the 6-hour RMSE values fall.

Overall, the year's average RMSE across the five stations for >1.0 ft surge was 0.40 ft and for >1.5 ft surge was 0.50 ft. JFK Airport has very little growth of RMSE with surge amplitude, suggesting that surges are fairly well-predicted and tides are the source of error at that station. RMSE for >1ft surge was 0.39 ft (0.11-0.76) and for >1.5 ft surge it was 0.45 ft (0.14-0.93).

Evaluating the uncertainty estimates (spread) of SFAS forecasts (**Figure 8**), there is a general trend between surge and COU, representing a tendency toward slight overestimation of uncertainty for negative surges and (at times) slight underestimation for positive surges. Cases with no surge have a 88-93% range, meaning that the spread of the 90% confidence interval on forecast water levels is a very good approximation of uncertainty. Cases with a negative surge have a COU from 84-94% and a slight tendency for uncertainty to be overestimated. Cases with a positive surge have a COU from 82-93% with a slight tendency for uncertainty to be underestimated.

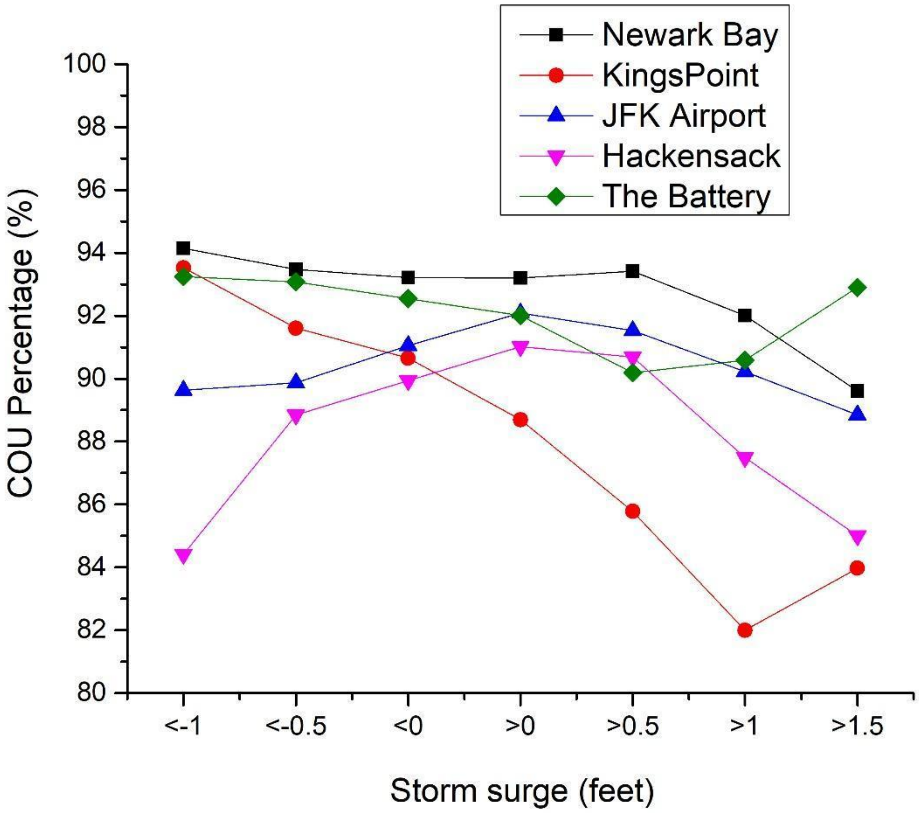

Figure 8: The Coverage of Observation by forecast area of Uncertainty with surge threshold for multi-stations comparison operated by the NYHOPS. The orange dashed line is the 90% threshold, which is the ideal value of COU because it is the uncertainty range provided with the forecasts.

5. Watch and Advisory Assessment

A component of SFAS that is being evaluated for the first time in 2021 is the Watch and Advisory email notification system. Users who have signed up on the SFAS webpage receive emails for specific stations of interest when a flood event is on the horizon or approaching within hours. Specifically, "Four-day" Flood Watches are triggered when the 95th percentile time series of a forecast period (+92 to 96h) exceeds the moderate flood threshold, and Flood Advisories are triggered when the central forecast time series (between 0 and 8h) exceeds the minor flood threshold (e.g. **Figure 2**).

An approach for evaluating the Watches and Advisories is to count "Hits", "Misses" and "False Alarms". Hits are cases where an email went out and an event did occur (e.g. moderate flood threshold was exceeded). Misses are cases where no email was sent but an event occurred. False Alarms are cases where an email was sent but no event occurred. A perfect score for Hits is 100%, for Misses is 0% and for False Alarms is 0% (Saleh et al. 2017). However, due to the random nature of coastal water levels, a probabilistic forecast system will never achieve these values and forecasters must seek a balance between maximizing Hits while not greatly increasing False Alarms (Verkade and Werner, 2011).

A few ratios of these counts are useful to consider relative to user interests: The Probability of Detection (POD), Hits / (Hits + Misses), and the False Alarm Rate (FAR), False / (Hits + False). A system should aim to maximize POD and keep FAR well below 100% and in tune with user needs and various costs and benefits (e.g. costs and benefits of actions taken pending a potential flood). If FAR is too high, then warning fatigue also becomes a potential problem and can lead to inaction.

Table 4: Results for SFAS Four-day Moderate Flood Watches

	Battery	**Kings Point**	**Hacken-sack**	**Newark Bay**	**JFK Airport**	**Harbor Totals**
Hit (Moderate flood)	0	2	0	0	0	2
Miss (Moderate flood)	0	0	0	0	0	0
False Alarm	1	8	19	4	1	33
POD	0%	100% (2/2)	0%	0%	0%	100% (2/2)
FAR	100% (1/1)	80% (8/10)	100% (19/19)	100% (4/4)	100% (1/1)	94% (33/35)

For 2021 Flood Watches across the five harbor stations, there were 2 Hits, 0 Misses and 33 False Alarms (**Table 4**). The POD across all stations was 100% and the FAR was 94%.

For 2021 Flood Advisories, based on cases where minor flooding was observed, there was a 63% POD and 56% FAR (**Table 5**). However, minor flooding (also known as nuisance flooding) rarely requires any response, so this low POD and low FAR may be acceptable for users. Evaluating the performance for cases where moderate flooding occurred and a warming was important due to more dangerous conditions, POD was 100% and FAR 98%, indicating a successful performance for warning but perhaps over-warning the public of impending moderate coastal flooding (**Table 6**).

Table 5: Results for SFAS Flood Advisories, relative to observed minor floods

	The Battery	**Kings Point**	**Hacken-sack**	**Newark Bay**	**JFK Airport**	**Harbor Totals**
Hit (minor flood)	2	20	17	1	6	46
Miss (minor flood)	3	13	2	1	8	27
False Alarm	3	6	48	0	2	59
POD	40% (2/5)	61% (20/33)	89% (17/19)	50% (1/2)	43% (6/14)	63% (46/73)
FAR	60% (3/5)	23% (6/26)	74% (48/65)	0% (0/1)	25% (2/8)	56% (59/105)

Table 6: Results for SFAS Flood Advisories, relative to observed moderate floods

	The Battery	Kings Point	Hackensack	Newark Bay	JFK Airport	Harbor Totals
Hit (Moderate flood)	0	2	0	0	0	2
Miss (Moderate flood)	0	0	0	0	0	0
False Alarm	5	24	65	1	8	103
POD	n/a	100% (2/2)	n/a	n/a	n/a	100% (2/2)
FAR	100% (5/5)	92% (24/26)	100% (65/65)	100% (1/1)	100% (8/8)	98% (103/105)

6. Discussion and Conclusions

The forecast accuracy during 2021 is slightly better than that during 2020. There are 4 major changes between the 2021 and 2020 forecast assessment that could affect the forecast accuracy: (1) We applied the new tide and tide bias correction data for the 2021 year operational forecast, which could provide improved tidal accuracy than 2020; (2) 2020 included more storm events than 2021, including the large storm surge event Isaias; (3) The 2020 assessment had about 3 months of relatively calm-weather forecast data missing; (4) Lastly, we stopped running most of the CMC ensemble members at the end of March in 2021 due to limitations of functioning HPC nodes and their being the worst performing members.

Forecast accuracy during the year's top five high water events was excellent, with RMSE of 0.80 ft for four-day lead times and 0.30 ft for those below one day. Performance across the full year's datasets was also excellent for a coastal total water level forecast system, with an average RMSE across five harbor stations of ~0.50 ft for four-day lead times and ~0.30 ft for those below one day. A major system improvement that occurred in 2017 is that forecasts have since extended out 105 hours, compared to 81 hours in 2015-6.

During storm surge events, the forecast accuracy results for 2021 were generally similar to those reported for winter 2015-2016 and slightly less than that reported for 2020, except this computation includes forecasts over the 1 day longer lead time. Across several harbor stations, SFAS then averaged a 0.6 feet RMSE for >1.33 ft surge (Georgas et al. 2016). For the five stations quantified in 2021, the average RMSE across the five stations in

Figure 7 for >1.0 ft surge was 0.40 ft and for >1.5 ft surge was 0.50 ft. Interpolating these results to match the 2016 result (>1.33 ft) gives a RMSE value of 0.46 ft, a smaller result. When comparing on an apples-to-apples basis (looking at harbor stations for the same range of lead times and winter months only), system accuracy in storm surge events during 2021 is slightly improved relative to 2015-2016. For cases with storm surge greater than 1.33 ft, the RMSE was 0.46 ft, whereas it was 0.50 ft for the similar in 2015-6 (Georgas et al. 2016). The improved accuracy is likely due to improvements in the system and its forcing data since 2016. These include improved CMC modeling and product resolution (improved from 1.0 degree to 0.5 degree resolution) and improved ECMWF-ENS and ECMWF-HRES modeling (but not product resolution - that is fixed with the price we pay for the products).

The uncertainty estimates (spread) of SFAS forecasts was very good, with harbor-average COU of ~90.3% for cases with small surges, compared with an ideal value of 90%. Harbor-average COU averaged ~91.0% for the larger negative surges, and ~88.1% for the larger positive surges. The range of COU including all five stations separately is from 82 to 94%, still close to 90% and indicating reasonable estimates of uncertainty. A similar NOAA ensemble water level forecast system that runs only with GEFS and CMC ensembles showed COU values for 2018-2019 storms from 30-60%, compared with an ideal value of 80% for that system (which provides an 80% uncertainty range), revealing that uncertainty estimates with that system were always too small in storm events (Liu and Taylor, 2020). Experimental runs (not operational) with the addition of ECMWF-ENS in the ensemble resulted in little or no improvement, on average (Liu and Taylor, 2020).

Over 1200 users have signed up for Flood Watches and Advisories since 2015, and this forecast assessment is the first to analyze the annual statistics of Flood Watches and Advisories. For 2021 moderate flood watches – The POD across all stations was 100% and the FAR was 94%. Given that these are based on the 95th percentile forecast exceeding the moderate flood threshold, it is unsurprising and by design that the FAR is high. The system previously provided 3-day Flood Watches, and these were turned off with the rationale that there were too many warnings. However, careful consideration should be given to whether there should be communications about an impending flood event between 4 days and 12 hours when the Flood Advisories begin.

For 2021 Flood Advisories, based on cases where minor flooding was observed, there was a 63% POD and 56% FAR. Based on cases where moderate flooding occurred, the POD was 100% and FAR was 98%. Given that the Advisories do not state what level of flooding may occur, and different locations have different thresholds for flooding (e.g. Orton et al. 2019), it is ambiguous how useful these results are to users. The SFAS website controls enable users to turn sites on and off and give them some measure of control for choosing whether the communications for a particular site is useful.

7. Future Recommendations

An important area of improvement is forcing model resolution, as discussed in detail in the 2020 forecast assessment. Isaias in 2020 was one case where resolution bias was a factor in reducing the accuracy and COU of the forecasts, as noted in the 2020 assessment and Ayyad et al. (2022). Nevertheless, recent analyses comparing NOAA's National Hurricane Center P-SURGE tropical cyclone storm surge forecast system shows that SFAS did considerably better for accuracy and spread during that event (Ayyad et al. 2022). Presently, limits on NOMADS data volumes and requests are preventing us from obtaining and using meteorological forcing products that are larger, higher-resolution datasets. Work is presently going toward both reducing NOMADS requests, data volumes, and also toward utilizing improved meteorological forcing datasets.

We recommend continued effort toward quantitative comparisons of Stevens and NOAA forecasts (P-ETSS, ET-SURGE, P-SURGE, ESTOFS). These are underway for the combined 2020-2021 data, with preliminary comparisons of P-ETSS and SFAS showing clearly that SFAS had better accuracy and spread. We also recommend continued assessment of different atmospheric ensemble and deterministic forecast products. This work can help determine improved weighting or changes to the ensemble of members that is running, as was done with omitting many CMC members in early 2021. Continued analysis of ensemble weighting will be beneficial, but preliminary results have not revealed simple answers nor a clear dependency on atmospheric forecast resolution.

While this report assesses statistics of Watch and Advisory communications, a user group would likely be needed to determine whether their settings should be adjusted to better address user needs. A future project could create a new website for SFAS and use the 1200+ user database for interviews or workshops to help define user interests.

For 2022 onward, a simpler approach will be used for evaluating past forecasts, where throughout the year, the 5th, 50th and 95th percentile time series are saved for all stations. This will enable a more comprehensive assessment of all stations, with less effort re-creating the ensemble forecasts from raw model outputs. However, it will result in data that is less useful for experimentation on processing methods. Given that we have sufficient data (two years) for the purposes of research on data processing methods, this new approach for assessment will be a step forward in breadth and ease.

For correspondence or questions, email Philip Orton at porton@stevens.edu

References

Ayyad, M., Orton, P.M., El Safty, H. and Hajj, M.R., 2022. Assessment of a Super-Ensemble Forecast for Coastal Storm Tide and Resurgence from Tropical Storm Isaias. In press, *Weather and Climate Extremes*.

Bruno, M.S., Blumberg, A.F. and Herrington, T.O., 2006. The urban ocean observatory-coastal ocean observations and forecasting in the New York Bight. In: Proceedings-Institute of Marine Engineering Science and Technology Part C, *Journal of Marine Science and Environment*, 4:31.

Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G. and Vitart, F. 2007, The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). *Q.J.R. Meteorol. Soc.*, 133: 681-695.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.*, 138, 1877–1901.

ECMWF, 2018. *Annual Report 2018*, Report, ECMWF. https://www.ecmwf.int/node/19127

Georgas, N., Orton, P., Blumberg, A., Cohen, L., Zarrilli, D. and Yin, L., 2014. The impact of tidal phase on Hurricane Sandy's flooding around New York City and Long Island Sound. *Journal of Extreme Events*, *1*(01), p.1450006.

Georgas, N. and Blumberg, A.F., 2010. Establishing confidence in marine forecast systems: The design and skill assessment of the New York Harbor Observation and Prediction System, version 3 (NYHOPS v3). In *Estuarine and Coastal Modeling (2009)* (pp. 660-685).

Georgas, N., Blumberg, A., Herrington, T., Wakeman, T., Saleh, F., Runnels, D., Jordi, A., Ying, K., Yin, L., Ramaswamy, V. and Yakubovskiy, A., 2016. The Stevens flood advisory system: Operational H3E flood forecasts for the greater New York/New Jersey metropolitan region. *Flood Risk Management and Response*, p.194.

Han, J. and Pan, H.L., 2011. Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Weather and Forecasting*, *26*(4), pp.520-533.

Jordi, A., Georgas, N., Blumberg, A., Yin, L., Chen, Z., Wang, Y., Schulte, J., Ramaswamy, V., Runnels, D. and Saleh, F., 2019. A next-generation coastal ocean operational system: Probabilistic flood forecasting at street scale. *Bulletin of the American Meteorological Society*, *100*(1), pp.41-54.

Liu, H., Taylor, A. and Kang, K., 2019, January. 3.8 LATEST DEVELOPMENT IN THE NWS'EXTRA-TROPICAL STORM SURGE MODEL, AND PROBABILISTIC EXTRA-TROPICAL STORM SURGE MODEL. In *99th American Meteorological Society Annual Meeting*.

Mukai, A.Y., Westerink, J., Luettich Jr, R.A., and Mark, D.. 2002. *Eastcoast 2001, a tidal constituent database for western North Atlantic, Gulf of Mexico, and Caribbean Sea*. DTIC Document.

Orton, P., Georgas, N., Blumberg, A. and Pullen, J., 2012. Detailed modeling of recent severe storm tides in estuaries of the New York City region. *Journal of Geophysical Research: Oceans*, *117*(C9).

Orton, P.M., Hall, T.M., Talke, S.A., Blumberg, A.F., Georgas, N. and Vinogradov, S., 2016. A validated tropical-extratropical flood hazard assessment for New York Harbor. *Journal of Geophysical Research: Oceans*, *121*(12), pp.8904-8929.

Orton, P. M., Chen, Z., El Safty, H., Ayyad, M., Datla, R., Miller, J., and Hajj, M.R. (2021), Stevens Flood Advisory System 2020 Ensemble Forecast Assessment: NY/NJ Harbor Area, 21 pp, Hoboken, New Jersey, USA.

Saleh, F., Ramaswamy, V., Wang, Y., Georgas, N., Blumberg, A., and Pullen, J. (2017), A multi-scale ensemble-based framework for forecasting compound coastal-riverine flooding: The Hackensack-Passaic watershed and Newark Bay, Advances in Water Resources, 110, 371-386.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., & Powers, J. G. 2005. A Description of the Advanced Research WRF Version 2 (No. NCAR/TN-468+STR). *University Corporation for Atmospheric Research*.

Taylor, P. K., and Yelland, M.J. 2001, The dependence of sea surface roughness on the height and steepness of the waves, J. Phys. Oceanogr., 31(2), 572-590.

Verkade, J.S. and Werner, M.G.F., 2011. Estimating the benefits of single value and probability forecasting for flood warning. *Hydrology and Earth System Sciences*, *15*(12), pp.3751-3765.